УДК 681.3.06
ББК 05.13.11

МОДЕЛЬ ИНФОРМАЦИОННОГО ПОИСКА НА ОСНОВЕ СЕМАНТИЧЕСКИХ МЕТАОПИСАНИЙ

Нгуен Ба Нгок[1], Тузовский А. Ф.[2]
(*Томский политехнический университет, Томск*)

Существующие подходы семантического поиска на основе онтологии, как правило, базируются на оценках концептуальной близости между элементами онтологии. В данной статье представлены методы вычисления оценки семантической близости между триплетами, а также между наборами триплетов. На основе этих оценок семантической близости представлена теоретическая модель семантического поиска, в которой документы и запросы представляются как наборы триплетов.

Ключевые слова: онтология, семантическая близость, семантический поиск, семантические метаописания, модель информационного поиска.

1. Введение

Под метаданными обычно понимаются данные о данных или информация об информации. Однако термин метаданные также используется по-разному в различных областях. В некоторых областях этот термин используется для обозначения машиночитаемой информации, одновременно в других областях термин *метаданные* используется для обозначения записи описания электронных ресурсов.

---

[1] *Нгуен Ба Нгок, аспирант* (*nguyen_bn@hotmail.com*).
[2] *Анатолий Федрович Тузовский, доктор технических наук, профессор* (*Томск, ул. Советская 84, тел.* (*3822*) *42-14-85*).

Важной причиной создания метаданных является возможность поиска информации с использованием релевантных критериев. Одним из примеров таких систем поиска по метаданным является сервис поиска файлов в операционной системе *Windows 7* (рис. 1), в котором для поиска файлов требуются задания значений трех их атрибутов: название, дата создания и размер.

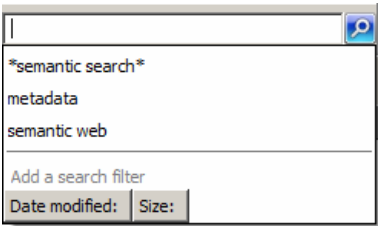

Рис. 1. Поисковый интерфейс в Windows

Информационный поиск по метаданным также широко применяется для поиска в электронных каталогах библиотек. На рис. 2 представлен интерфейс поиска научно-технической библиотеки ТПУ (http://www.lib.tpu.ru/catalog_arm.html – дата обращения 11.04.2012).

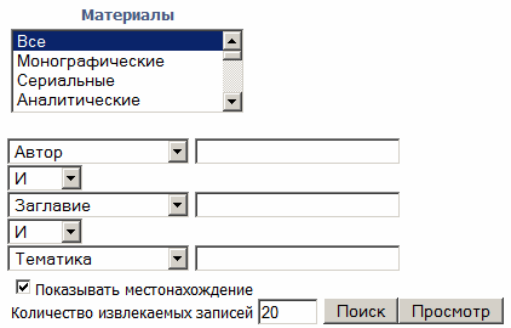

Рис. 2. Интерфейс поиска по метаданным

С использованием онтологии в работе [2] предложен метод создания особого вида метаданных – метаописаний документов, представляющих собой наборы простых высказываний вида

«субъект–предикат–объект», которые также называются триплетами и отражают основные семантики описываемых документов.

Отмечается, что такие метаописания являются ценными источниками информации для выполнения поиска и с их применением возможно значительное улучшение функциональности поисковых систем. В связи с этим в данной работе представлено описание модели информационного поиска на основе таких метаописаний, которое является теоретическим обоснованием для последующих программных реализаций. Предлагаемая модель представлена на рис. 3.

Известно, что основными функциями поисковых систем являются: 1) индексирование; 2) формирование запросов; 3) обработка запросов. В предлагаемой модели семантического поиска *задача индексирования* (1) подразделяется на две подзадачи: аннотирование и индексирование созданных метаданных.

Задача *аннотирования* заключается в создании семантических метаописаний для имеющихся документов. Семантические метаописания создаются с использованием терминологии онтологии предметной области, которая редактируется с помощью редактора (рис. 3), и могут быть разделены на контекстные и контентные метаданные, которые соответственно описывают контексты и контенты (содержания) объектов-документов. При этом метаописания документов могут быть сформированы без участия человека (автоматическим способом) либо с участием человека (полуавтоматическим либо ручным путем) [2].

Задача индексирования метаданных заключается в сборе метаописаний документов в базах данных системы (которые также называются индексами) для цели эффективного выполнения запросов.

Формирование поисковых запросов (2): с помощью поискового интерфейса системы клиенты формируют свои информационные потребности в виде множеств триплетов – запросы. Формированные запросы затем направляются в подсистему обработки запросов для выполнения.

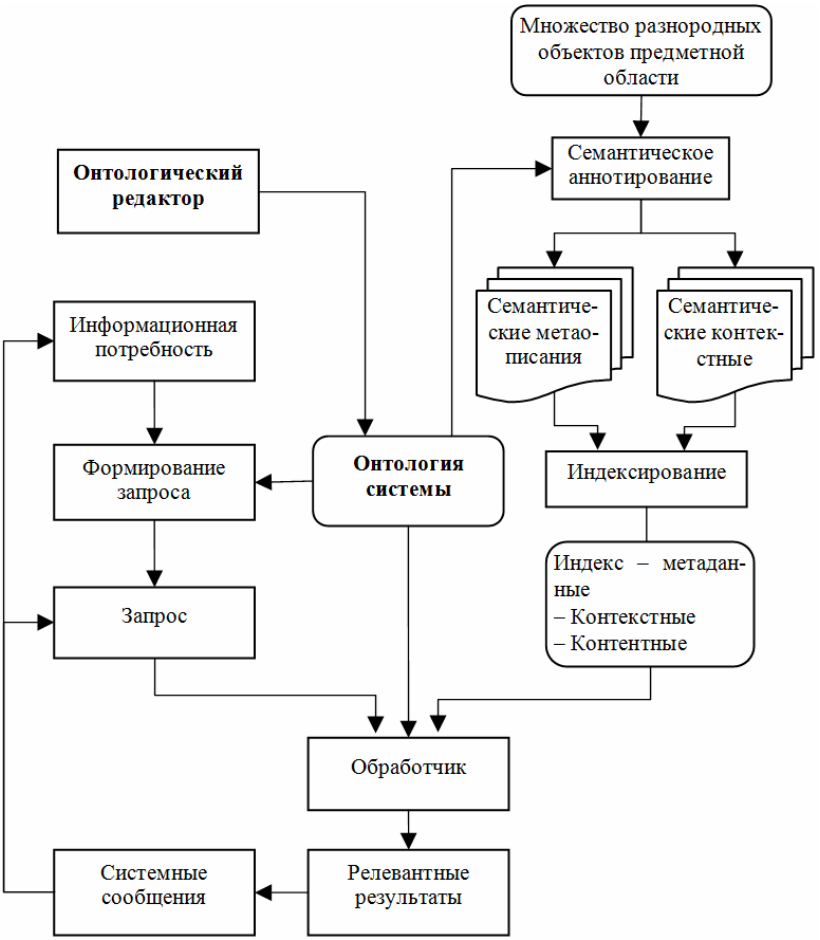

Рис. 3. Предлагаемая модель информационного поиска

Обработка запросов (3): данная задача заключается в сравнении запросов пользователей (клиентов) с метаописаниями документов индекса. При этом онтология используется как фундаментальный компонент для вычисления оценки семантической близости между ними.

Затем на основе вычисленных оценок близости определяется множество релевантных документов и формируются систем-

ные сообщения, которые затем возвращаются клиентам. При анализе полученных результатов в некоторых случаях уточняются информационные потребности пользователей. В результате этого переформируются поисковые запросы.

Подробные описания проблемы индексирования и структуры индексов представлены в [17]. Далее в данной работе представляется описание предлагаемой модели поиска и алгоритма обработки запросов.

2. Постановка задачи

Определения понятия семантики, семантической близости, и семантического поиска были представлены в работе [7] следующим образом: под семантикой текста обычно понимается его «смысл», который автор текста хотел передать посредством символов, однако для компьютерной системы смысл текста строго зависит от контекста, где он определяется и обрабатывается. В области семантического веба лучшим средством представления семантики является онтология.

Оценкой близости между документом и запросом является числовое значение, которое выражает степень сходства между ними; оценка близости называется оценкой семантической близости, если и только если она определена на основе семантики документов и запросов.

Далее подходом семантического поиска называется подход поиска, в котором используется концепция семантической близости для сопоставления документов заданному запросу.

В данной работе используются следующие упрощенные обозначения компонентов онтологии: онтология O представляет собой знаковую систему $O = <C, E, R, T>$, где C – множество понятий (классов); E – множество экземпляров понятий; R – множество предикатов – типов отношений; T – множество отношений, которые задают следующие виды связи между сущностями:

1. Частичный порядок на множествах C и R, задающий отношения *is-a* – «подкласс-суперкласс» – «выше-ниже».

2. Отношение между понятиями, которое представляет собой триплет вида $<c_1 - r_1 - c_2>$, где $c_1, c_2 \in C$; $r_1 \in R$.
3. Отношение между экземплярами, которое представляет собой триплет вида $<e_1 - r_1 - e_2>$, где $e_1, e_2 \in E$; $r_1 \in R$.
4. Отношение между предикатами, которое представляет собой триплет вида $<r_1 - r_i - r_2>$, где $r_1, r_2, r_i \in R$.

Допустим:

1) На основе онтологии предметной области O для каждого документа d_i коллекции документов $D = \{d_i\}$ созданы семантические метаописания

$$m(d_i) = \{t_1, t_2, \ldots, t_{n(i)}\},$$

где $n(i)$ – количество триплетов в логическом представлении документа d_i; t_i – *RDF*-триплеты – кортежи вида $\langle s_i, p_i, o_i \rangle$, где s_i и o_i включены в объединение C_i и E_i, а p_i включен в R.

2) Каждый запрос q, данный пользователем из множества запросов Q, также состоит из множества триплетов

$$q = \{t_1, t_2, \ldots, t_{n(q)}\},$$

где $n(q)$ – количество триплетов, содержащихся в запросе q.

3) Определена весовая функция w, которая определяет значимость любого триплета $t \in T$ (T – множество возможных триплетов) при описании документов d_i и запроса q:

$0 \le w(t, d_i) \le 1$, где $t \in T$, $d_i \in D$,

$0 \le w(t, q) \le 1$, где $t \in T$, $q \in Q$.

Решаемая задача заключается в том, что для каждого запроса q требуется определить подмножество *RES* множества документов D, которое состоит из релевантных документов для заданного запроса q – результирующее множество. Документ d_i считается релевантным заданному запросу q, если и только если оценка семантической близости между ними превышает некоторую пороговую величину. При этом для вычисления близости между документом и запросом используются их семантические метаописания.

3. Модель информационного поиска на основе нечетких множеств

В связи с необходимостью выражения разных степеней значимости триплетов (их весовые коэффициенты) и релевантности между документами и запросами для цели ранжирования результатов, в данной работе используется модель нечеткого множества для описания предлагаемого алгоритма поиска.

3.1. НЕЧЕТКОЕ МНОЖЕСТВО: БАЗОВЫЕ ОПРЕДЕЛЕНИЯ

Определение: четкое множество A определяется характеристической функцией fA, которая назначает каждому элементу u универсального множества U единственное значение из множества $\{0, 1\}$ следующим образом:

$fA: U \to \{0, 1\}$,

и

$$fA(u) = \begin{cases} 1, \text{если } u \in A, \\ 0, \text{если иначе.} \end{cases}$$

Функция fA различает элементы множества A от тех элементов, которые не принадлежат данному множеству [22].

Нечеткое множество A определяется характеристической функцией μA, которая ставит каждому элементу u универсального множества U единственное значение интервала $[0, 1]$:

$\mu_A: U \to [0, 1]$,

где μ_A выражает степени членства элементов нечеткого множества A следующим образом: $\mu_A(u) = 1$ означает, что u принадлежит A полностью; $0 < \mu_A(u) < 1$ – что u частично принадлежит A; $\mu_A(u) = 0$ – что u полностью не принадлежит A.

Исходя из этих определений видно, что характеристическая функция четкого множества является частным случаем характеристической функции нечеткого множества, которая имеет только два значения 0 и 1.

Нечеткое подмножество A универсального множества U, определенное характеристической функцией μA, обозначается следующим образом:

$$A = \{\mu_A(u)/u \mid u \in U\},$$

или

$$A = \mu_A(u_1)/u_1 + \mu_A(u_2)/u_2 + \cdots + \mu_A(u_n)/u_n = \sum_{i=1}^{n} \mu_A(u_i)/u_i,$$

где + и $\sum$ означают операцию объединения; n – количество элементов множества U.

3.2. НЕЧЕТКОЕ МНОЖЕСТВО: БАЗОВЫЕ ОПЕРАЦИИ

Размер нечеткого подмножества конечного универсального множества U является скалярной величиной, которая равна сумме степеней членства его элементов. Данная величина вычисляется по следующей формуле:

$$|A| = \sum_{u \in U} \mu_A(u).$$

Результатом операции α-срезки нечеткого множества A для порогового значения $\alpha \in [0, 1]$ является следующее четкое множество:

$${}^{\alpha}A = \{x \mid \mu_A(x) \geq \alpha\}.$$

Результаты операций пересечения и объединения двух нечетких множеств A и B определяются соответственно с помощью функций треугольной нормы (t-норма) и треугольной конормы (t-конорма):

$$\mu_{A \cap B}(u) = T(\mu_A(u), \mu_B(u)),$$

$$\mu_{A \cup B}(u) = S(\mu_A(u), \mu_B(u)),$$

где T – функция *t-нормы*; S – функция *t-конормы*. При этом имеются следующие популярные определения функции *t-нормы* и *t-конормы*:

1. Минимум T_M и максимум S_M:
   $T_M(a,b) = \min(a,b)$; $S_M(a,b) = \max(a,b)$.
2. Вероятностное произведение T_P и вероятностная сумма S_P:
   $T_P(a.b) = a \cdot b$; $S_P(a,b) = a + b - a \cdot b$.
3. t-норма Лукасевича T_L и t-конорма Лукасевича S_L:
   $T_L(a,b) = \max(a + b - 1, 0)$; $S_L(a,b) = \min(a + b, 1)$.
4. Сильное произведение T_D и сильная сумма S_D:

$$T_D(a,b)=\begin{cases} a, \text{если } b=1, \\ b, \text{если } a=1, \\ 0, \text{если иначе,} \end{cases} \quad ; \quad S_D(a,b)=\begin{cases} a, \text{если } b=0, \\ b, \text{если } a=0, \\ 1, \text{если иначе.} \end{cases}$$

3.3. НЕЧЕТКОЕ МНОЖЕСТВО: ФОРМАЛИЗАЦИЯ ПРЕДЛАГАЕМОЙ МОДЕЛИ ПОИСКА

На основе семантических метаописаний документов (наборы триплетов) и весовой функции имеется следующее определение нечеткого подмножества I множества пар *документ–триплет* – множество индекса:

$$I=\{\mu_I(d,t)/(d,t) \mid d \in D; t \in T\},$$

где $\mu_I(d, t) = w(t, d)$ – значимости триплета t при описании документа d.

На основе индекса I семантические метаописания документа $d \in D$ могут быть легко преобразованы в нечеткие подмножества множества триплетов (его логическое представление в предлагаемой модели информационного поиска) следующим образом:

$$I_d=\{\mu_{I(d)}(t)/t \mid t \in T\}, \text{ где } \mu_{I(d)}(t)=\mu_I(d,t).$$

Подобно метаописаниям документов запрос q также может быть представлен как нечеткое подмножество множества триплетов следующим образом:

$$I_q=\{\mu_{I(q)}(t)/t \mid t \in T\}, \text{ где } \mu_{I(q)}(t)=w(t,q).$$

В предлагаемой модели информационного поиска функция семантической близости определяет для каждого запроса q следующее нечеткое подмножество релевантных документов множества документов D:

$$RES=\{\mu_{RES}(d)/d \mid d \in D\},$$

где $\mu_{RES}(d) = sim_{sem}(d, q)$ – семантическая близость между документом d и запросом q. При этом для сравнения документов и запросов используются их семантические метаописания (логические представления), т.е.

$$sim_{sem}(d,q)=sim_{sem}(I_d, I_q).$$

Для фильтрации документов с низкими степенями близости используется пороговое значение $\alpha \in [0, 1]$ по следующему правилу: *документ d считается релевантным запросу q, если и только если оценка близости между ними превышает пороговое значение α*. Соответственно нечеткое подмножество результатов определяется следующим образом:

$$RES_\alpha = \{\mu_{RES}(d)/d \mid d \in {}^{\alpha}RES\},$$

где ${}^{\alpha}RES$ – результирующее множество операции α-срезки для множества результатов *RES*.

Далее описываются методики вычисления различных видов оценок семантической близости на основе онтологии.

4. Вычисление семантической близости

Учет структуры онтологии и семантики отношений позволяет вычислять оценки семантической близости между элементами онтологии (понятия, экземпляры, связи – предикаты). Эти оценки близости называются элементарными оценками близости, на основе которых определяются близости между триплетами. Оценки близости между триплетами затем используются для определения близости между метаописаниями.

Исходя из описанной выше структуры триплетов (см. раздел 2), для их сравнения требуются вычисления следующих видов элементарных оценок близости между: 1) понятиями; 2) понятиями и экземплярами; 3) экземплярами; 4) предикатами. Методы вычисления каждого типа элементарной оценки близости представляются далее.

В данной работе используется упрощенное определение триплетов, т.е. не учитывается случай, когда субъекты триплетов являются литеральными значениями. С учетом этого помимо элементарных оценок семантической близости требуется определение близости между строковыми литеральными значениями. Данная оценка близости не является семантической и может быть вычислена с использованием, например, расстояния Левенштейна [25].

Далее в разделах 4.1 и 4.2 представлены обзоры подходов семантической близости между понятиями. Данная оценка близости также известна как концептуальная близость. Представляемый обзор является расширением работы [5], т.е. приведены более подробные описания рассматриваемых подходов и добавлены новые подходы по концептуальной близости.

4.1. КОНЦЕПТУАЛЬНАЯ БЛИЗОСТЬ В ТАКСОНОМИИ

В этом разделе представлен обзор подходов вычисления концептуальной близости на основе таксономии, т.е. учитывается только семантическое отношение «выше-ниже» (*is-a* – также известно как отношение «родитель–ребенок»). Рассматриваемые подходы группируются согласно свойствам, которые используются для вычисления близости. При этом чаще всего используются следующие характеристики понятий: 1) длина пути между понятиями; 2) глубина в таксономии; 3) информационное содержание; 4) множество родительских понятий.

Кроме того, в разделе *гибридные подходы* рассматриваются методы комбинирования разных мер близости для получения более эффективного измерения близости.

Подходы на основе длины пути.

Rada R., *Mili H.*, *Bicknell E. и Blettner M.* В [31] предложено определение семантического расстояния между понятиями (обратная величина близости, т.е. чем больше расстояние, тем меньше близость и обратно) как количество ребер пути между ними в таксономии:

$$dist_{Rada}(c_1, c_2) = \min(|path(c_1, c_2)|),$$

где $|path(c_1, c_2)|$ – количество ребер пути от c_1 до c_2. Пути между понятиями определены с учетом таксономии как неориентированный граф.

Hirst G., *St-Onge D.* В [20] представлена мера близости, по которой ограничиваются характеристики путей между понятиями. При этом учитываются только те пути, которые содержат не больше 5 ребер или соответствуют одному из 8 шаблонов, представленных в [20]. Близость по допустимому пути вычисляется следующим образом:

$$sim_{Hirst\&St-Onge}(c_1, c_2) = S - \text{длина пути} - k \cdot \text{кол. изм. напр.}$$

Следовательно, чем длиннее путь и больше количество изменений по направлению, тем меньше близость между понятиями.

Лукашевич Н.В., *Добров Б.В*. Подобно подходу *Hirst & St-Onge* в [6] ограничивается конфигурация путей, используемых при вычислении близости. При этом рассматриваются пути, состоящие из совокупности иерархических отношений (выше–ниже – *is-a*, часть – целое – *partOf*, и несимметричная ассоциация) либо направленные в одну сторону, либо включающие ровно один перегиб (изменение по направлению).

Bulskov H., *Knappe R.*, *Andreasen T.* В работе [15] близость между понятиями x и y вычисляется как максимальное произведение весовых коэффициентов ребер путей между ними. По этому методу для отношения *is-a* задаются два параметра *gen*, *spec* $\in [0, 1]$, которые соответственно выражают близости в направлении обобщения и детализации.

Для пути $P = \{c_1, c_2, \ldots, c_n\}$ между понятиями x и y, где $c_1 = x$, $c_n = y$, определяются следующие характеристики:

$g(P) = |\{i \mid c_i - isa - c_{i+1}\}|$ – количество ребер в направлении обобщения;

$s(P) = |\{i \mid c_{i+1} - isa - c_i\}|$ – количество ребер в направлении детализации.

С учетом множественного наследования, допустим, что $\{P_1, P_2, \ldots, P_m\}$ – множество возможных путей между понятиями x и y. Тогда близость между ними определяется следующим образом:

$$sim_{WSP}(x, y) = \max_{i=1\ldots m}(spec^{s(Pi)} \cdot gen^{g(Pi)}).$$

Подходы на основе длины пути и глубины вершин.

Sussna M. В работе [36] представлена мера семантического расстояния между понятиями *WordNet* (обратная величина близости). При этом семантическое расстояние между любыми понятиями c_1 и c_2 вычисляется как сумма расстояний между соседними понятиями, входящими в пути между ними.

Семантическое расстояние между понятиями c_1 и c_2, связанное отношением r, вычисляется по формуле

$$dist_{sussna}(c_1, c_2) = \frac{\omega(c_1 \overset{r}{\rightarrow} c_2) + \omega(c_2 \overset{r^*}{\rightarrow} c_1)}{2d},$$

где r^* –обратное отношение отношения r; весовые коэффициенты определяются по следующей формуле:

$$\omega(c_1 \overset{r}{\rightarrow} c_2) = \max_r - \frac{\max_r - \min_r}{n_r(c_1)},$$

где d – глубина ребра таксономии – максимальное значение глубины двух понятий; $\max_r$, $\min_r$ – максимальное и минимальное значение весового коэффициента отношения r; $n_r(X)$ – количество выходов из понятия X отношения r.

Вычисление семантического расстояния в таксономии является частным случаем оригинального алгоритма, т.е. учитываются только отношение «выше-ниже» (*hypernymy* – обобщение) и его обратное отношение – отношение детализации (*hyponymy*). Согласно работе [36] весовые коэффициенты этих отношений варьируются в диапазоне от 1 до 2.

Wu Z., *Palmer M.* В работе [39] концептуальная близость между понятиями c_1 и c_2 определена следующим образом:

$$sim_{Wu\&Palmer}(c_1, c_2) = \frac{2N_3}{N_1 + N_2 + 2N_3}.$$

Пусть c_3 – ближайший общий родитель понятий c_1 и c_2, тогда имеются следующие определения параметров заданной формулы: N_1 – количество вершин пути от c_1 до c_3 – длина пути между ними; N_2 – количество вершин пути от c_2 до c_3; N_3 – количество вершин пути от c_3 до коренного понятия таксономии – её глубины.

Leacock C., *Chodorow M.* В работе [24] оценка семантической близости между понятиями вычисляется по следующей формуле:

$$sim_{Leacock\&Chodorow}(c_1, c_2) = -\log\left(\frac{Np(c_1, c_2)}{2D}\right),$$

где $Np(c_1, c_2)$ – длина кратчайшего пути между ними (количество вершин); D – максимальная глубина таксономии.

Nguyen H.A. В [29] предложено измерение семантического расстояния, которое является функцией двух параметров – длины кратчайшего пути между вершинами и общей специфичности двух вершин. Вводится понятие общей специфичности двух вершин *CSpec*:

$$CSpec(c_1, c_2) = N - N(LCS(c_1, c_2)),$$

где N – максимальная глубина таксономического дерева. Чем меньше специфичность двух вершин, тем большей информацией они обладают и близость их больше.

Семантическое расстояние между понятиями c_1 и c_2 определяется по следующей формуле:

$$SemDist(c_1, c_2) = \log((d(c_1, c_2) - 1)^{\alpha} \cdot (CSpec(c_1, c_2))^{\beta} + k),$$

где $\alpha > 0$, $\beta > 0$; $k \geq 1$ – константы (обеспечивают нелинейность и положительность функции *SemDist*); $d(c_1, c_2)$ – длины кратчайшего пути между ними.

Haase P., *Siebes R.*, *Harmelen F.* В [19] представлена следующая мера близости между понятиями c_1 и c_2:

$$sim_{Haas}(c_1, c_2) = \begin{cases} e^{-\alpha l} \cdot \dfrac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}}, \text{если } c_1 \neq c_2, \\ 1, \text{если иначе.} \end{cases}$$

где l – длина кратчайшего пути между понятиями c_1 и c_2 и h – глубина ближайшего общего родителя понятий c_1 и c_2.

Параметры $\alpha, \beta \geq 0$ соответственно регулируют влияние расстояния между понятиями и глубины ближайшего общего родителя при вычислении близости.

Как было определено в [19], оптимальными значениями параметров являются $\alpha = 0{,}2$ и $\beta = 0{,}6$.

Подходы на основе информационного содержания

Resnik P. В работе [32] предложено следующее определение близости между понятиями c_1 и c_2:

$$sim_{\text{Resnik}}(c_1, c_2) = \max_{c \in S(c1, c2)} (IC(c)),$$

где $S(c_1, c_2)$ – множество ближайших верхних общих понятий для c_1 и c_2; $IC(c)$ – информационное содержание понятия c. При

этом имеется следующее определение величины информационного содержания понятия таксономии:

пусть C – множество понятий таксономии и для каждого понятия $c \in C$ определена вероятность $p(c)$ того, что встречается экземпляр понятия c в обучающей коллекции текстов. В работе [34] информационное содержание понятия c определено следующим образом:

$$IC(c) = -\log p(c).$$

В работе [32] предложено, что встречаемость каждого термина обучающей текстовой коллекции учитывается в подсчете частоты тех понятий таксономий, которые включают данный термин. Исходя из данного правила, частота понятия в коллекции определяется следующим образом:

$$freq(c) = \sum_{t \in words(c)} freq(t),$$

где $words(c)$ – множество терминов, которые по смыслу принадлежат понятию c. Принадлежность термина понятию определяется ручным путем и учитывается наследование между понятиями, т.е. если термин t принадлежит понятию x *и* x *ISA* y, то термин t также принадлежит понятию y. По примеру таксономии, которая представлена на рис. 4, при встрече существительных *dime* увеличиваются частоты понятия *dime*, *coin*, *cash* и т.д.

Вероятность того, что встречается понятие c, определяется по следующей формуле:

$$p(c) = \frac{freq(c)}{N},$$

где N – количество аннотированных терминов обучающей коллекции.

Jiang J., *Conrath D.* В [21] представлена мера близости, которая является функцией двух параметров: количество ребер пути между понятиями и информационные содержания понятий. При этом информационные содержания используются как поправочные величины.

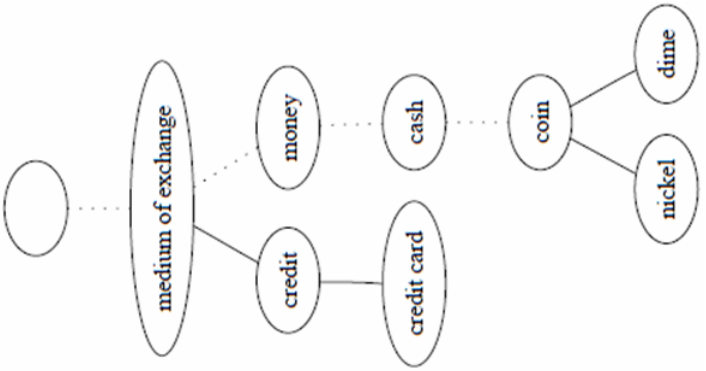

Рис. 4. Фрагмент таксономии WordNet

Общая формула для определения весового коэффициента ребра между понятием c и родительским понятием p представляется следующим образом:

$$wt(c,p)=\left(\beta+(1-\beta)\frac{E^*}{E(p)}\right)\cdot\left(\frac{d(c)+1}{d(p)}\right)^{\alpha} LS(c,p)T(c,p),$$

где $d(c)$ – глубина понятия c; $E(p)$ – количество детей понятия p (локальная плотность); E^* – среднее значение плотности, определенной в таксономии; $LS(c,p)$ – сила связи между c и p; $T(c,p)$ – коэффициент типа отношения; параметры $\alpha \geq 0$ и $0 \leq \beta \leq 1$ регулируют соответственно глубину понятия и плотность понятия. По результатам экспериментов, которые представлены в [21], имеются следующие оптимальные значения для этих параметров: $\alpha_{\text{опт}} = 0{,}5$; $\beta_{\text{опт}} = 0{,}3$. В таксономии, где используется единственное отношение *is-a*, коэффициент типа отношения $T(c,p) = 1$ и не влияет на вычисление расстояния между вершинами.

В работе [21] отмечено, что мощность отношения $LS(c,p)$ пропорциональна вероятности $p(c|p)$ того, что встречается экземпляр понятия c при условии, что встречается экземпляр его родительского понятия p:

$$LS(c,p) = -\log p(c \mid p).$$

Данная вероятность определена по следующей формуле:

$$p(c \mid p) = \frac{p(c \cap p)}{p(p)}.$$

Исходя из способа вычисления вероятности, который представлен в работе [32], получается, что $p(c \cap p) = p(c)$, так как любой экземпляр понятия c является экземпляром родительского понятия p. Следовательно,

$$p(c \mid p) = \frac{p(c)}{p(p)} \quad \text{и} \quad LS(c, p) = IC(c) - IC(p),$$

где $IC(c)$ и $IC(p)$ – информационное содержание понятий c и p соответственно. В работе [21] семантическое расстояние между понятиями определено как сумма весов ребер, входящих в кратчайший путь между ними:

$$dist_{Jiang\&Cpmrath}(c_1, c_2) = \sum_{c \in (path(c1,c2) - LSuper(c1,c2)} wt(c, parent(c)),$$

где $path(c_1, c_2)$ – множество вершин пути, соединяющего c_1 и c_2; $LSuper(c_1, c_2)$ – множество общих родителей понятий c_1 и c_2; $parent(c)$ – родительское понятие понятия c.

В частном случае, когда учитываются только силы связей, имеется следующее упрощенное определение семантического расстояния:

$$Dist(c_1, c_2) = IC(c_1) + IC(c_2) - 2IC(LSuper(c_1, c_2)).$$

Lin D. В работе [26] предложено следующее определение близости между понятиями таксономии:

$$sim_{Lin}(c_1, c_2) = \frac{2 \log p(LUB(c_1, c_2))}{\log p(c_1) + \log p(c_2)},$$

где $LUB(c_1, c_2)$ – верхняя грань понятий c_1 и c_2; $p(c_1)$, $p(c_2)$ – информационное содержание понятий c_1 и c_2, которые могут быть определены на основе обучающей коллекции по методу, который представлен выше.

Подходы на основе множества родительских понятий

Тузовский А.Ф. В [11, 12] представлено следующее определение понятия множества родительских понятий C_{ANC}: «*для каждого понятия $c_i \in C$ существует множество $C_{ANC}(c_i)$, являющееся подмножеством C и содержащее понятия, пред-*

шествующие понятию c_i, а также само понятие c_i.», где C – множество имеющихся понятий.

Для оценки семантической близости двух понятий c_k и c_l вводятся два показателя, основанные на сравнении множеств $C_{ANC}(c_i)$:

$$sim(c_k, c_l) = k_{st} \cdot \frac{|C_{ANC}(c_k) \cap C_{ANC}(c_l)|}{|C_{ANC}(c_k) \cup C_{ANC}(c_l)|},$$

где $k_{st} = \begin{cases} 1, \text{если } C_{ANC}(c_k) \cup C_{ANC}(c_l) = C_{ANC}(c_l), \\ 0, \text{если иначе.} \end{cases}$

Гибридные подходы

В [5] представлено следующее определение гибридных подходов: *гибридные меры являются свертками некоторых мер близости понятий. Чем полнее будут учитываться характеристики двух сущностей с разных точек зрения, тем более качественную меру близости можно получить. В связи с этим наиболее перспективными представляются именно гибридные меры, сочетающие несколько подходов.*

Чаще всего в гибридных мерах используется аддитивная свертка:

$$S(c_1, c_2) = \sum_{i=1}^{n} \omega_i \cdot sim^i(c_1, c_2),$$

где sim^i – i-я мера близости; вес ω_i определяет важность данной меры близости; сумма весов равна единице; n – количество мер близости.

Распространенная модификация аддитивной свертки основана на сигмоидальной функции, которая позволяет повысить веса мер, имеющих большие значения и практически пренебречь мерами с малыми значениями:

$$sig(x) = \frac{1}{1 + e^{-\alpha x}}, \text{ где } \alpha > 0;$$

$$S(c_1, c_2) = \sum_{i=1}^{n} \omega_i sig(sim^i(c_1, c_2)).$$

4.2. КОНЦЕПТУАЛЬНАЯ БЛИЗОСТЬ В ОНТОЛОГИИ

Вышеописанные методы вычисления концептуальной близости базируются на отношении «выше-ниже». Однако нет необходимости ограничить измерения близости в использовании только этого отношения. В онтологии количество возможных семантических отношений неограниченно, поэтому семантическая близость в онтологии является более широким понятием, чем в таксономии.

Оценка концептуальной близости в онтологии может означать доли общих частей сущностей и степень связанности между ними. Например, понятия «электрический автомобиль» и «автомобиль» являются близкими понятиями, однако «бензин» и «автомобиль» являются связанными понятиями, и степень связанности между ними может быть выражена оценкой близости в онтологии. Проблема вычисления оценки семантической близости с учетом различных типов семантических отношений рассмотрена в работах [13, 16, 23, 37].

Castano S., *Ferrara A.*, *Montanelli S.*, *Racca G.* В [16] представлена симметричная мера семантической близости, в которой учитываются различные семантические отношения. При этом для каждого семантического отношения задается вес, который принимает значение в диапазоне [0, 1]. Семантическая близость заданного пути вычисляется как произведение весов её ребер.

Допустим, что $\{P_1, P_2, \ldots, P_k\}$ – множество возможных путей между c_1 и c_2, тогда близость между ними вычисляется следующим образом:

$$sim_{Cast}(c_1, c_2) = \begin{cases} \max_{i=1..k}(w(P_i)), \text{если } k > 0, \\ 0, \text{если иначе.} \end{cases}$$

где k – количество путей между c_1 и c_2; $w(P_i)$ – оценка близости на основе пути P_i:

$$w(P_i) = \prod_{j=1..n} \omega_{i,j},$$

где n – количество ребер пути P_i; ω_{ij} – весовой коэффициент j-го ребра пути P_i.

Tversky A. В работе [37] представлена мера близости, которая является основой для многих современных подходов вычисления близости в онтологии. По этому методу близость между понятиями a и b является функцией трех аргументов $A \cap B$, $A - B$, $B - A$, где A, B – множество свойств этих понятий. Эта функция должна удовлетворять аксиомам монотонности, независимости, разрешимости и инвариантности и определяется формулой (*contrast model*):

$$sim(a,b) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A),$$

В развитие модели Тверски была построена *ratio model*:

$$sim(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A - B) + \beta f(B - A)}.$$

В большинстве методов вычисления близости используется *ratio model*, а в качестве функции f – мощность множества – аргумента.

Knappe R. В [23] представлена мера близости на основе контекстных множеств, которые определяются следующим образом: *контекстное множество заданного понятия состоит из тех понятий онтологии, которые достигнуты из данного понятия. Понятие y считается достигнутым из понятия x, если и только если существует хотя бы один путь между ними*.

Схема алгоритма определения контекстного множества A заданного понятия c_1 представлена на рис. 4. По данному алгоритму, сначала $A = \{c_1\}$. Затем контекстное множество A расширяется по следующей формуле:

$$extend(A) = A \cup \{y \mid x \in A \wedge y \notin A \wedge r(x,y) \in T\},$$

где $r \in R^*$ – множество допущенных семантических отношений; T – множество связей между элементами онтологии. Процесс расширения повторяется итерационно до тех пор, пока не достигнуто условие завершения.

В качестве условия завершения можно использовать, например, ограничение по количеству операции расширения или по признаку добавления новых элементов после выполнения операции.

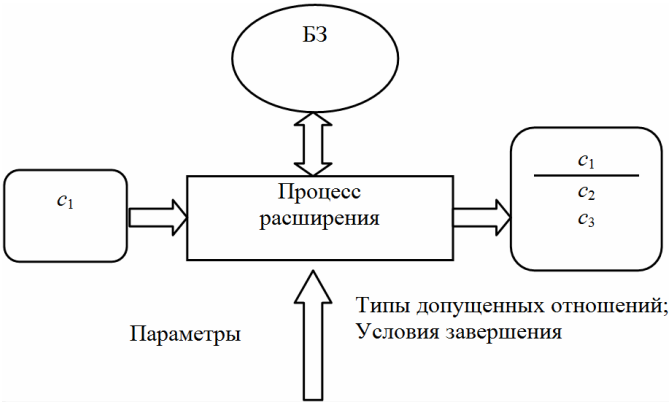

Рис. 5. Определение контекстного множества вершин

На основе контекстных множеств в [23] близость между понятиями c_1 и c_2 вычисляется следующим образом:

$$sim_{sharednodes}(c_1, c_2) = \rho \cdot \frac{|A(c_1) \cap A(c_2)|}{|A(c_1)|} + (1-\rho) \cdot \frac{|A(c_1) \cap A(c_2)|}{|A(c_2)|},$$

где $\rho \in [0, 1]$ – параметр для объединения двух составляющих.

Andreasen T., *Knappe R.*, *Bulskov H.* Как отмечено в работе [13], по предыдущему методу [23] значимости вершины для вычисления близости неодинаковые. Поэтому имеется возможность улучшения данного метода путем определения весовых коэффициентов для элементов контекстного множества.

С учетом весов элементов контекстное множество понятия c_1 есть нечеткое подмножество μA множества понятий и экземпляров онтологии:

$$\mu A = \omega_1 / c_1 + \omega_2 / c_2 + \ldots + \omega_n / c_n,$$

при этом весовые коэффициенты соответствуют степени принадлежности элементов нечеткому множеству.

Схема модифицированного алгоритма определения контекстного множества для понятия c_1 с учетом весовых коэффициентов представлена на рис. 6.

Сначала контекстное множество вершины $\mu A = \{\omega_1 / c_1\}$, где $\omega_1 = 1$ – весовой коэффициент понятия c_1. Затем контекстное множество μA расширяется идентично случаю описанного выше

метода [23]. При этом весовой коэффициент новой вершины определяется как произведение весовых коэффициентов семантического отношения и исходной вершины.

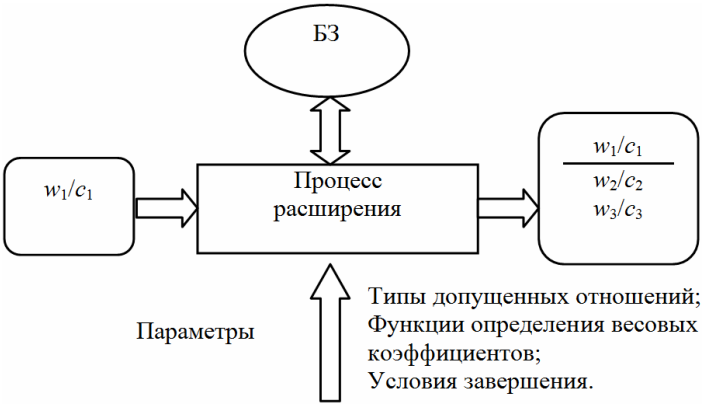

Рис. 6. Алгоритм определения контекстного множества вершин с учетом весовых коэффициентов

Допустим, что новая вершина y добавляется с использованием связи $r(x, y)$ и для отношения r определен весовой коэффициент $f(r)$. Тогда весовой коэффициент вершины y определяется следующим образом:

$$\omega_y = f(r) \cdot \omega_x ,$$

где ω_x – весовой коэффициент вершины x.

Весовые коэффициенты семантических отношений являются важными параметрами алгоритма и могут быть настроены по-разному в разных предметных областях. Метод автоматического определения оптимальных коэффициентов с использованием обучающей выборки представлен далее в разделе 4.7.

4.3. БЛИЗОСТЬ МЕЖДУ ЭКЗЕМПЛЯРАМИ

Тузовский А.Ф. В [11] предложена мера близости, согласно которой оценка семантической близости $sim_I(i_1, i_2)$ двух экземпляров i_1 и i_2 складывается из их реляционной близости и близости их типов (понятия, к которым они относятся).

Отмечается, что в онтологии O для каждого экземпляра $i_x \in I$ существуют:

1) непустое множество $C_{INST}(i_x)$, включающее понятия, к которым относится экземпляр i_x:

$$C_{INST}(i_x) \neq \varnothing, \quad C_{INST}(i_x) = \{c_j \in C \mid is(i_x, c_j)\};$$

2) множество $R_{INST}(i_x)$, включающее все конкретизированные отношения экземпляра i_x:

$$R_{INST}(i_x) = \{r_i \in R \mid r_i(i_x, i_y)\}.$$

При этом реляционная близость $sim_{IL}(i_1, i_2)$ позволяет оценить схожесть двух экземпляров исходя из их отношений с другими экземплярами онтологии:

$$sim_{IL}(i_1, i_2) =$$

$$= \begin{cases} \dfrac{|R_{EQU}(i_1, i_2)|}{|R_{INST}(i_1) \cap R_{INST}(i_2)|}, \text{если } R_{INST}(i_1) \cap R_{INST}(i_2) \neq \varnothing, \\ 0, \text{если иначе,} \end{cases}$$

где $R_{EQU}(i_1, i_2) = \{r_i \in R \mid r_i(i_1, i_z) \wedge r_i(i_2, i_z)\}$.

Близость типов определяется следующим образом:

$$sim_{IC}(i_1, i_2) = \frac{\sum\limits_{c_1 \in C_{INST}(i_1)} \max\limits_{c_2 \in C_{INST}(i_2)} (sim(c_1, c_2))}{n},$$

где n – размер множества $C_{INST}(i_1)$; $sim(c_1, c_2)$ – близость понятий c_1, c_2, которая может быть вычислена с использованием вышеописанных методов концептуальной близости.

В итоге семантическая близость двух экземпляров определяется как аддитивная свертка двух оценок близости:

$$sim_I(i_x, i_y) = k_{IC} \cdot sim_{IC}(i_x, i_y) + k_{IL} \cdot sim_{IL}(i_x, i_y),$$

где k_{IC}, $k_{IL} \in [0, 1]$; $k_{IL} = 1 - k_{IC}$.

Коэффициенты k_{IC} и k_{IL} позволяют настраивать процедуру вычисления семантической близости двух экземпляров. Если экземпляры описаны в онтологии в основном с помощью связей с другими экземплярами или конкретными значениями, то необходимо установить соотношение $k_{IC} < k_{IL}$. В противном случае необходимо установить соотношение $k_{IC} > k_{IL}$.

4.4. СЕМАНТИЧЕСКАЯ БЛИЗОСТЬ МЕЖДУ ПРЕДИКАТАМИ

Как описано выше в условии задачи, предикаты организуются в таксономии предикатов. Поэтому вышеописанные меры близости в таксономии являются подходящими для вычисления близости между ними.

Однако учет только отношения *is-a* будет недостаточным, так как в онтологии близости между предикатами также выражаются явным образом через отношение *sameAs* (идентичность) и *invertOf* (инверсное отношение).

Семантическая близость между предикатами, связанными отношением *sameAs*, очевидна, принимает максимальное значение близости и равна единице. В случае если предикаты связаны отношением *invertOf*, предполагается, что близость между ними равна –1. Следовательно, расширяется диапазон значения близости между предикатами. Близость между предикатами принимается в диапазоне [–1; 1]. При этом значение близости между предикатами < 0 означает *инверсное сравнение триплетов*.

Предполагаемая мера близости для предикатов с учетом этих особенностей представляется в разделе 4.6, а инверсное сравнение триплетов – в разделе 4.8.

4.5. СЕМАНТИЧЕСКАЯ БЛИЗОСТЬ РАЗНОТИПНЫХ ЭЛЕМЕНТОВ

В работе [11] представлены методы определения близости между разнотипными элементами, согласно которым сравнение двух разнотипных элементов онтологии возможно лишь с некоторым допущением, которое выражается соответствующим коэффициентом:

Для вычисления семантической близости между *понятием и экземпляром* используется коэффициент $d_{CI} \in (0, 1]$, который выражает близость между родительским понятием и экземпляром:

$$sim_{CI}(c,i) = d_{CI} \cdot \max_{c_x \in C_{INST}(i)}(sim(c,c_x)).$$

Аналогично, для вычисления семантической близости между экземпляром и понятием используется коэффициент

$d_{IC} \in (0, 1]$, который выражает близость между экземпляром и соответствующим родительским понятием:

$$sim_{IC}(i, c) = d_{IC} \cdot \max_{c_x \in C_{INST}(i)}(sim(c_x, c)).$$

4.6. ОБОБЩЕНИЕ ПОДХОДА ВЗВЕШЕННОГО КРАТЧАЙШЕГО ПУТИ

В этом разделе представляется модификация меры близости, представленной в работе [15], с учетом всех видов семантических отношений, которая должна быть подходящей для вычисления всех видов близости между компонентами триплетов.

По предлагаемой модификации:

1. Для отношения «родитель–ребенок» (*is-a*) задаются два коэффициента *gen* и *spec*, которые соответственно выражают близость в направлении обобщения и детализации.

2. Для отношения *instanceOf* (связывает понятие с экземплярами понятий) задаются два параметра d_{IC}, $d_{CI} \in [0, 1]$, которые соответственно выражают близость экземпляра понятию и близость понятия экземпляру.

3. Коэффициенты близости для отношения *sameAs* (синонимы) и *invertOf* (обратные отношения) соответственно равны 1 и –1.

4. Для остальных семантических отношений r_i определяется весовой коэффициент *wr*, который выражает семантическую близость по этим отношениям.

Для P – путь между сущностями x и y (которые могут быть понятиями, экземплярами, или предикатами) определяются следующие характеристики:

1. $s(P)$ – количество ребер в направлении детализации;
2. $g(P)$ – количество ребер в направлении обобщения;
3. $ic(P)$ – количество ребер от экземпляра до понятия;
4. $ci(P)$ – количество ребер от понятия до экземпляра;
5. $inv(P)$ – количество ребер инверсного отношения;
6. $oth(P)$ – количество ребер остальных отношений.

Оценка близости между сущностями x и y согласно пути P определяется по следующей формуле:

$$sim^{P}_{GWSP}(x, y) =$$
$$= (-1)^{inv(P)} \cdot spec^{s(P)} \cdot gen^{g(P)} \cdot d_{IC}{}^{ic(P)} \cdot d_{CI}{}^{ci(P)} \cdot wr^{oth(P)}.$$

Допустим, что $PATH = (P_1, P_2, \ldots, P_k)$ является множеством всех возможных путей между сущностями x и y, тогда близость между ними определяется следующим образом:

$$sim_{GWSP}(x, y) = sim^{P_{\max}}_{GWSP}(x, y),$$

где путь $P_{\max}$ с максимальной оценкой близости определяется по следующему условию:

$$| sim^{P_{\max}}_{GWSP}(x, y) | = \max_{P \in PATH} (| sim^{P}_{GWSP}(x, y) |).$$

В частном случае, если $PATH = \varnothing$, то значение близости считается равным нулю:

$$sim_{GWSP}(x, y) = 0, \text{ если } PATH = \varnothing.$$

4.7. НАСТРОЙКИ КОЭФФИЦИЕНТОВ

Параметры алгоритма могут быть настроены ручным или автоматическим путем. В литературе представлены методы автоматической настройки оптимальных значений коэффициентов с помощью обучаемой нейронной сети [18] или генетического алгоритма [35]. В этом разделе представлен примитивный способ определения приблизительных оптимальных значений коэффициентов по методу максимизации оценки меры близости.

Основная идея предлагаемого метода заключается в том, что чем больше значение коэффициента корреляции между оценками близости алгоритма и оценками близости экспертов, тем эффективнее считается алгоритм, и при оптимальных значениях параметров алгоритма значение коэффициента корреляции является максимальной, так как согласно [28, 33] оценка близости имеет субъективный характер.

По предлагаемому методу область значения $[a, b]$ коэффициента k сначала дискретизируется на $n + 1$ точек:

$$k_i = a + i \cdot \frac{b - a}{n},$$

где $i = 0, 1, \ldots, n$.

Затем оптимальные значения коэффициентов определяются методом полного перебора по следующему условию:

$$corr(k_{i(\max)}, \ldots, k_{m(\max)}) = \max(corr),$$

где $corr(k_{1(i)}, \ldots, k_{m(i)})$ – значение коэффициента корреляции меры близости с использованием заданных параметров; $\max(corr)$ – максимальное значение коэффициента корреляции для рассматриваемой меры близости.

Для вычисления значения коэффициента корреляции меры близости сначала вычисляются близости между сущностями обучающей коллекции с помощью алгоритма. Затем вычисляется значение коэффициента корреляции с оценками близости экспертов по известной формуле Пирсона.

Значение коэффициента корреляции Пирсона между сериями n измерений случайных переменных X и Y, обозначенных как x_i и y_i, где $i = 1, 2, \ldots, n$, вычисляется следующим образом:

$$r = \frac{n\sum x_i y_i + \sum x_i \cdot \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2} \cdot \sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

В итоге для нахождения оптимальных значений параметров требуется выполнение n^m итераций, где n – частота дискретизации; m – количество параметров. Соответственно, по предлагаемому методу, чем больше частота дискретизации, тем выше точность определения оптимальных значений параметров. Однако увеличивается требуемое время для работы алгоритма.

4.8. СЕМАНТИЧЕСКАЯ БЛИЗОСТЬ МЕЖДУ ТРИПЛЕТАМИ

В [11, 12] близость между триплетами вычисляется как среднее значение оценок близости между их компонентами:

$$sim(< t_1, k_1 >, < t_2, k_2 >) =$$

$$= \frac{sim(s_1, s_2) + sim(p_1, p_2) + sim(o_1, o_2)}{3} \cdot f(k_1, k_2).$$

В [30] близость между триплетами $t_1 = (s_1, p_1, o_1)$ и $t_2 = (s_2, p_2, o_2)$ вычисляется следующим образом:

Если $sim(s_1, s_2) + sim(o_1, o_2) \geq sim(s_1, o_2) + sim(o_1, s_2)$, то близость между t_1 и t_2 определяется следующим образом:

$$sim(t_1, t_2) = \omega_{ss} \cdot sim(s_1, s_2) + \omega_{pp} \cdot sim(p_1, p_2) + \omega_{oo} \cdot sim(o_1, o_2).$$

Иначе используется следующая формула:

$$sim(t_1, t_2) = \omega_{so} \cdot (sim(s_1, o_2) + sim(s_2, o_1)) + \omega_{pp} \cdot sim(p_1, p_2),$$

где ω_{ss}, ω_{pp}, ω_{oo}, ω_{so} – весовые коэффициенты, которые выражают значимости составляющих оценок семантической близости и

$$\omega_{ss} + \omega_{pp} + \omega_{oo} = \omega_{pp} + 2 \cdot \omega_{so} = 1.$$

Представленные методы вычисления семантической близости между триплетами неверны в случае, когда триплеты выражают противоречивые мнения о связи между одинаковыми сущностями, например две фразы «команда А–выиграть–команда Б» и «команда А–проиграть–команда Б». С использованием методов суммирования получается высокая близость между этими высказываниями, хотя они имеют противоречивые смыслы.

Кроме того, если p_1 и p_2 – обратные отношения, то триплет $t_2 = < s_2, p_2, o_2 >$ имеет такое же значение, как триплет $t_2^* = < o_2, p_1, s_2 >$. Следовательно, имеется следующее равенство:

$$sim_{sem}(t_1, t_2) = sim_{sem}(t_1, t_2^*).$$

В общем случае, если путь между предикатами в онтологии содержит нечеткое количество отношения *invertOf* (инверсное отношение), то при вычислении близости между триплетами предлагается использование *инверсного сравнения* триплетов, т.е. вычисляются близости между субъектами и объектами (в отличие от *прямого сравнения*, когда субъект сравнивается с субъектом и объект сравнивается с объектом).

Исходя из указанных предположений, в данной работе предлагается следующее определение близости между триплетами $t_1 = <s_1, p_1, o_1>$ и $t_2 = < s_2, p_2, o_2 >$:

$$sim_{sem}(t_1, t_2) = \begin{cases} |k| \cdot \dfrac{sim_{sem}(s_1, s_2) + sim_{sem}(o_1, o_2)}{2}, & \text{если } k > 0, \\ |k| \cdot \dfrac{sim_{sem}(s_1, o_2) + sim_{sem}(o_1, s_2)}{2}, & \text{если иначе,} \end{cases}$$

где $k = sim_{sem}(p_1, p_2)$ – оценка близости между предикатами.

4.9. СЕМАНТИЧЕСКАЯ БЛИЗОСТЬ МЕЖДУ МЕТАОПИСАНИЯМИ

В литературе представлены различные подходы вычисления близости между метаописаниями, которые состоят из множеств триплетов [1, 3–4, 9–12, 30, 38]. При анализе существующих подходов выделяются три направления решения данной задачи (количественной оценки близости между метаописаниями). По *первому направлению* метаописания рассматриваются как пакет триплетов – *BOT* (***B**ag **O**f **T**riples*). При этом элементарными единицами сравнения являются триплеты. По *второму* – метаописания рассматриваются как *RDF*-граф, вершинами которого являются субъекты и объекты триплетов, а ребрами – семантические отношения. При этом элементарными единицами сравнения являются элементы онтологии. По *последнему направлению* метаописания сначала преобразуются в аналогичные форматы – пакет слов – *BOW* (***B**ag **O**f **W**ords*), или пакет понятий *BOC* (***B**ag **O**f **C**oncepts*), затем применяются известные меры близости для этих форматов.

К *первому направлению* относятся подходы, представленные в работах [11, 12, 30]. В [11, 12] оценка семантической близости между метаописаниями I_d и I_q вычисляется как сумма близостей между всеми парами триплетов следующим образом:

$$sim(I_d, I_q) = \sum_{t_i \in I_q} \sum_{t_j \in I_q} sim(t_i, t_j).$$

В [30] семантическая близость между метаописаниями вычисляется по следующей формуле:

$$sim(I_d, I_q) = \frac{1}{|I_d| + |I_q|} \cdot \left(\sum_{t_i \in I_d} sim(t_i, I_q) + \sum_{t_j \in I_q} sim(t_j, I_d) \right),$$

где близость между триплетом и множеством триплетов есть максимальная близость между данным триплетом и триплетами множества.

С другой стороны, формат *BOT* является аналогом формата пакета слов (*BOW*) – логическое представление текстов в классических системах информационного поиска. Поэтому если учитывать триплеты как ключевые слова, то можно применять

классические меры близости для вычисления близости между метаописаниями.

Например, в [10] для вычисления весов триплетов в *BOT* предлагается использовать меру *TF*/*IDF*, рассматривая каждый триплет как отдельное «слово», а для вычисления близости – косинусную меру.

К *второму направлению* относятся подходы, представленные в работах [1, 4, 38].

В [38] предлагается вычислять семантическую близость между метаописаниями, представленными в формате *BOT*, на основе соответствия *RDF*-графов. Проблема определения близости графов в такой формулировке лежит в рамках известной задачи проверки изоморфизма графов [3], которая является *NP*-полной. Однако учет специфики *RDF*-графов позволяет избежать проблемы *NP*-полноты. Приведен полиномиальный алгоритм поиска наилучшего отображения, при котором сумма близостей соответствующих вершин и дуг максимальна.

В [1, 4] близость двух *RDF*-графов вычисляется как взвешенная сумма близости вершин и близости по отношениям. Близость вершин оценивается долей совпадающих вершин с учетом весов, близость по отношениям – суммой долей общих дуг отдельно по каждому типу отношений с учетом весов отношений.

Главная идея *последнего направления* заключается в том, что метаописания сначала могут быть преобразованы в пакеты понятия (*BOC*), которые состоят из элементов триплетов, или в пакеты ключевых слов (*BOW*), которые состоят из меток элементов триплетов. Затем для вычисления близости между метаописаниями можно применить известные меры близости для форматов *BOC* или *BOW* соответственно.

При таком преобразовании потеряются явные семантические связи между сущностями. Из-за чего снижается точность выполнения запроса. Однако повышается производительность обработки, так как форматы *BOC* и *BOW* являются более простыми форматами по сравнению с форматом *BOT*.

Далее в этом разделе предлагаются два метода вычисления близости между метаописаниями, которые относятся к первому направлению: 1) метод суммирования, являющийся модификацией метода, представленного в [11, 12]; 2) метод максимального паросочетания в двудольном взвешенном графе.

Метод суммирования оценок близости, который представлен в [11, 12], не учитывает количество триплетов метаописаний документа и запроса и количество их общих триплетов. С учетом этих характеристик представляется следующее определение близости между метаописаниями I_d и I_q:

$$sim_{sem}(I_d, I_q) = \frac{|I_d \cap I_q|}{\max(|I_{d(i)}|)} \cdot \frac{I_d \cdot I_q}{|I_q| \cdot |I_q|},$$

где $|I_d|$, $|I_q|$ и $|I_d \cap I_q|$ соответственно являются размерами представления документа d, представления запроса q и их пересечения; $\max(|I_{d(i)}|)$ – константа – максимальный размер метаописаний коллекции документов; I_d, I_q – сумма оценок близости между триплетами попарно, которая определяется следующим образом:

$$I_d \cdot I_q = \sum_{x \in T, y \in T} \mu_{I(d)}(x) \cdot \mu_{I(q)}(y) \cdot sim_{sem}(x, y).$$

Основная идея метода *максимального паросочетания во взвешенном двудольном графе* заключается в сопоставлении каждого триплета запроса с единственном триплетом метаописания документа таким образом, чтобы сумма близости между ними являлась максимальной. При этом находятся оптимальные отображения между триплетами метаописаний.

Двудольный граф является особенным графом, обладающим следующими двумя свойствами: 1) все вершины двудольного графа могут быть распределены в двух непересекающихся множествах (левое и правое множество вершин); 2) любое ребро графа соединяет только вершину из левого множества с вершиной правого множества.

По предлагаемому алгоритму сначала строится двудольный взвешенный граф $BG = <V, E>$, где множество вершин графа V

является объединением левого множества вершин и правого множества вершин:

$$V = V_L \cup V_R ,$$

где $V_L = {}^{\alpha}I_d$ – левое множество вершин, которое состоит из триплетов метаописания документа, и $V_R = {}^{\alpha}I_q$ – правое множество вершин, которое состоит из триплетов метаописания запроса.

Для каждой пары вершин $v_{t(i)} \in V_L$ и $v_{t(j)} \in V_R$ определяется ребро графа ($<v_{t(i)}, v_{t(j)}>$, e_{ij}) с весовым коэффициентом e_{ij}, который равен оценке семантической близости между соответствующими триплетами:

$$e_{ij} = e(v_{t(i)}, v_{t(j)}) = \mu_{I(d)}(t_i) \cdot \mu_{I(q)}(t_j) \cdot sim_{sem}(t_i, t_j) .$$

После определения двудольного графа определяется его максимальное паросочетание. При этом паросочетание двудольного графа является множеством пар вершин графа вида

$$P = \{(v_{l(i)}, v_{r(i)}) \mid v_{l(i)} \in V_L ; v_{r(i)} \in V_R ;\ i = 1,..., \min(|V_L|, |V_R|)\} ,$$

в котором каждая вершина встречается только один раз.

Максимальным является паросочетание, имеющее максимальную сумму весовых коэффициентов ребер:

$$sum_{BG}(P_{\max}) = \max_{P \in P^*}(sum_{BG}(P)) ,$$

где P^* – множество всех возможных паросочетаний; $sum_{BG}(P)$ – сумма весовых коэффициентов ребер паросочетания P графа BG:

$$sum_{BG}(P) = \sum_{(v_{l(i)}, v_{r(i)}) \in P} e(v_{l(i)}, v_{r(i)}) .$$

Эффективным алгоритмом для нахождения максимального паросочетания во взвешенном двудольном графе является алгоритм *Hungarian* [14] (алгоритм был предложен венгерским математиком *Egervary*). В итоге оценка семантической близости между документом d и запросом q определяется как сумма весовых коэффициентов ребер максимального паросочетания, деленная на нормирующий множитель:

$$sim_{sem}(d, q) = sim_{sem}(I_d, I_q) = \frac{sum_{BG}(P_{\max})}{\max(|I_{d(i)}|)} .$$

где $\max(|I_{d(i)}|)$ – нормирующий множитель – максимальный размер метаописаний коллекции документов.

5. Выполнение запроса

В качестве примера рассматривается информационная поисковая система со следующими исходными данными:

множество триплетов T, которое состоит из трех триплетов:

$T = \{t_1, t_2, t_3\}$,

и следующая коллекция документов D:

$D = \{d_1, d_2, d_3\}$.

Опускаются подробности структуры онтологии; допустим, имеются следующие метаописания документов:

$m(d_1) = \{t_1, t_2\}$; $m(d_2) = \{t_1, t_3\}$; $m(d_3) = \{t_2, t_3\}$.

При этом каждый триплет t_i имеет вид $<s_i, p_i, o_i>$, т.е.

$t_1 =< s_1, p_1, o_1 >$, $t_2 =< s_2, p_2, o_2 >$, $t_3 =< s_3, p_3, o_3 >$.

Без учета подробности процесса вычисления близости допустим, что имеется следующая функция базовых оценок семантической близости:

$sim_{sem}(s_3, s_1) = 0{,}9$; $sim_{sem}(p_3, p_1) = 1{,}0$; $sim_{sem}(o_3, o_1) = 0{,}7$;

$sim_{sem}(s_3, s_2) = 0{,}6$; $sim_{sem}(p_3, p_2) = 1{,}0$; $sim_{sem}(o_3, o_2) = 0{,}6$;

$sim_{sem}(s_2, s_1) = 0{,}3$; $sim_{sem}(p_2, p_1) = 1{,}0$; $sim_{sem}(o_2, o_1) = 0{,}3$;

$sim_{sem}(x, x) = 1{,}0$;

$sim_{sem}(x, y) = 0{,}0$ – в остальных случаях, когда $x \neq y$,

на основе которых определяется следующая функция семантической близости между триплетами:

$$sim_{sem}(t_3, t_1) = 1 \cdot \frac{0{,}9 + 0{,}7}{2} = 0{,}8;$$

$$sim_{sem}(t_3, t_2) = 1 \cdot \frac{0{,}6 + 0{,}6}{2} = 0{,}6;$$

$$sim_{sem}(t_2, t_1) = 1 \cdot \frac{0{,}3 + 0{,}3}{2} = 0{,}3;\ sim_{sem}(t, t) = 1{,}0;$$

$sim_{sem}(x, y) = 0{,}0$ – в остальных случаях, если $x \neq y$.

Для весовых функций:

$$w(t,d)=\begin{cases}1, \text{если } t\in d,\\ 0, \text{если иначе,}\end{cases}$$

и

$$w(t,q)=\begin{cases}1, \text{ если } t\in q,\\ 0, \text{ если иначе.}\end{cases}$$

Имеется следующее нечеткое множество индекса:

$$I=\{1/(d_1,t_1)+1/(d_1,t_2)+0/(d_1,t_3)+$$
$$+1/(d_2,t_1)+0/(d_2,t_2)+1/(d_2,t_3)+$$
$$+0/(d_3,t_1)+1/(d_3,t_2)+1/(d_3,t_3)\},$$

и соответствующие представления документов:

$$I_{d(1)}=\{1/t_1+1/t_2+0/t_3\};\quad I_{d(2)}=\{1/t_1+0/t_2+1/t_3\};$$
$$I_{d(3)}=\{0/t_1+1/t_2+1/t_3\}.$$

Для заданного запроса $q = \{t_1, t_2\}$ оценки семантической близости между документами и запросом **по первому методу** вычисляются следующим образом:

$$sim_{sem}(d_1,q)=\frac{2}{2}\cdot\frac{1\cdot1\cdot1+1\cdot1\cdot0{,}3+1\cdot1\cdot0+1\cdot1\cdot1}{2\cdot2}=0{,}6;$$

$$sim_{sem}(d_2,q)=\frac{1}{2}\cdot\frac{1\cdot1\cdot1+1\cdot1\cdot0{,}8+1\cdot1\cdot0+1\cdot1\cdot0{,}6}{2\cdot2}=0{,}3;$$

$$sim_{sem}(d_3,q)=\frac{1}{2}\cdot\frac{1\cdot1\cdot0{,}3+1\cdot1\cdot0{,}8+1\cdot1\cdot1+1\cdot1\cdot0{,}6}{2\cdot2}=0{,}35.$$

На основе вычисленных оценок семантической близости определяется следующее множество результатов:

$$RES=0{,}6/d_1+0{,}35/d_3+0{,}3/d_2,$$

при этом знак «+» означает операцию объединения элементов в множество.

По второму методу. Сначала строятся следующие взвешенные двудольные графы документов и запроса (рис. 7):

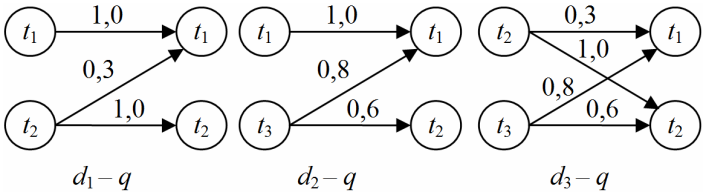

Рис. 7. Взвешенные двудольные графы

На основе полученных взвешенных двудольных графов имеются следующие оценки семантической близости документов запросу:

$$sim_{sem}(d_1, q) = \frac{1{,}0 + 1{,}0}{2} = 1{,}0;$$

$$sim_{sem}(d_2, q) = \frac{1{,}0 + 0{,}6}{2} = 0{,}8;$$

$$sim_{sem}(d_3, q) = \frac{1{,}0 + 0{,}8}{2} = 0{,}9.$$

Множество результатов в данном случае представляется следующим образом:

$$RES = 1{,}0 / d_1 + 0{,}9 / d_3 + 0{,}8 / d_2.$$

Исходя из представленного примера видно, что по данным алгоритмам для определения множества результатов требуется вычисление оценок близости всех документов. Из-за этого в случае обработки большой коллекции документов без применения метода оптимизации процесса вычисления возможно требуется много времени для выполнения запросов, и следовательно снижается эффективность системы.

6. Оптимизация выполнения запроса

В этом разделе представлены два метода повышения скорости обработки запросов: фильтрация коллекции документов с помощью инвертированного индекса и применение статических оценок близости. Эти методы могут быть применены отдельно или совместно в зависимости от доступных ресурсов вычислительных систем.

Первый способ оптимизации выполнения запроса заключается в определении нечеткого подмножества D_q множества документов D, документы из которого могут быть релевантными запросу q. Затем сравнения выполняются только в данном множестве D_q.

Данное множество D_q называется контекстным множеством запроса q. В данной статье предлагается определение контекстного множества D_q как объединение списков релевантных документов следующим образом:

$$D_q = \bigcup_{t \in q} I_t ,$$

где I_t – список релевантных документов триплета t. На основе индекса I список I_t определяется следующим образом:

$$I_t = \{\mu_{I(t)}(d) / d \mid d \in D\} ,$$

где

$$\mu_{I(t)}(d) = \mu_I(d, t) .$$

С использованием контекстного множества D_q имеется следующее оптимизированное определение множества результатов запроса q:

$$RES = \{\mu_{RES}(d) / d \mid d \in {}^{\alpha} D_q\} ,$$

где ${}^{\alpha}D_q$ – четкое множество, полученное в результате операции α-срезки над нечетком множеством D_q.

При этом ускорение получается за счет того, что размер множества ${}^{\alpha}D_q$ меньше, чем размер коллекции документов D. Однако снижается полнота результатов в связи с возможными ошибками фильтрации.

По второму методу скорость выполнения запроса может быть увеличена за счет использования статических оценок близости между элементами метаописаний. Статический метод означает, что оценки семантической близости вычисляются заранее до выполнения запроса и сохраняются в виде двухмерного массива.

В предлагаемой модели поиска данная идея применима для вычисления элементарных оценок близости и близости между

триплетами. На основе этих оценок возможно определить следующие нечеткие множества:

1) нечеткие подмножества множества пар элементов-триплетов:

$$S_A = \{\mu_{S(A)}(x, y)/(x, y) \mid x, y \in C_i \cup E_i\},$$

где $\mu_{S(A)}(x, y) = sim_{sem}(x, y)$.

$$S_P = \{\mu_{S(P)}(x, y)/(x, y) \mid x, y \in P_i\},$$

где $\mu_{S(P)}(x, y) = sim_{sem}(x, y)$.

2) нечеткое подмножество множества пар триплетов:

$$S_T = \{\mu_{S(T)}(x, y)/(x, y) \mid x, y \in T\},$$

где $\mu_{S(T)}(x, y) = sim_{sem}(x, y)$.

Для данного метода необходимо большое дисковое пространство для хранения массива оценок близости, так как требуется определение оценки семантической близости попарно. Примитивным методом решения этой проблемы является удаление малых элементов массивов, т.е. сохраняются только те пары, близость которых выше, чем заданное пороговое значение α.

7. Выводы

В описанной модели семантического поиска элементарными единицами для составления поисковых запросов и метаописаний документов являются триплеты.

В работе предложены общий метод для вычисления близости между компонентами триплетов (см. раздел 4.6), метод вычисления близости между триплетами (см. раздел 4.8) и две схемы вычисления семантической близости между метаописаниями: 1) сумма близости между составляющими триплетами; 2) максимальное паросочетание во взвешенном двудольном графе (см. раздел 4.9).

Описаны два метода оптимизации обработки запроса: фильтрация исходной коллекции документов с помощью инвертированного индекса и использование статических оценок близости. Отличительная особенность структуры инвертированного индекса, используемая в первом методе, заключается в

использовании триплетов для составления словаря указателей. Поэтому для реализации такого типа инвертированных файлов требуется эффективный метод организации словарей триплетов, исследование данной проблемы представлено в работе [8].

Второй метод заключается в том, что при вычислении семантической близости документа запросу составляющие оценки семантической близости могут быть вычислены статическим (когда оценки близости вычислены заранее и сохранены на диске) либо динамическим способом (когда оценки близости вычисляются во время выполнения запроса). По сравнению с динамическим методом статический метод увеличивает производительность системы, однако требуются большее дисковое пространство и дополнительные вычислительные затраты при индексировании.

Литература

1. БОГАТЫРЕВ М.Ю., ЛАТОВ В.Е., СТОЛБОВСКАЯ И.А. *Применение концептуальных графов в системах поддержки электронных библиотек* // Труды 9-й Всероссийской науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – Переславль, 2007. – Т. 2. – С. 104–110.
2. ГУБИН М.Ю., РАЗИН В.В., ТУЗОВСКИЙ А.Ф. *Методы создания семантических метаописаний документов с применением семантических сетей, фреймовых моделей и частотных характеристик* // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2010. – Т. 2, №2. – С. 227–229.
3. ГЭРИ М., ДЖОНСОН Д. *Вычислительные машины и труднорешаемые задачи*. – М.: Мир, 1982. – 192 с.
4. КАРПЕНКО А.П. *Оценка релевантности документов онтологической базы знаний* // Электронное научно-техническое издание «Наука и образование». – URL: http://technomag.edu.ru/doc/157379.html (дата обращения: 23.07.2012).

5. КРЮКОВ К.В., ПАНКОВА Л.А., ПРОНИНА В.А., ШИПИЛИНА Л.Б. *Меры семантической близости в онтологиях* // Проблемы управления. – 2010. – №2. – С. 2–14.
6. ЛУКАШЕВИЧ Н.В., ДОБРОВ Б.В. *Тезаурус русского языка для автоматической обработки больших текстовых коллекций* // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002 / Под ред. А.С. Нариньяни. – М.: Наука, 2002. – Т. 2. – С. 338–346.
7. НГУЕН Б.Н., ТУЗОВСКИЙ А.Ф. *Обзор подходов семантического поиска* // Доклады Томского государственного университета систем управления и радиоэлектроники. – 2010. – Т. 2, №2. – С. 234–237.
8. НГУЕН Б.Н., ТУЗОВСКИЙ А.Ф. *Оптимизация хранения словаря триплетов с использованием числовых идентификаторов* // Научно-технический вестник Поволжья. – 2012. – №2. – С. 235–245.
9. ПАНКОВА Л.А., ПРОНИНА В.А., КРЮКОВ К.В. *Онтологические модели поиска экспертов в системах управления знаниями научных организаций* // Проблемы управления. – 2011. – №6. – С. 52–60.
10. РАБЧЕВСКИЙ Е.А. *Автоматическое построение онтологий на основе лексико-синтаксических шаблонов для информационного поиска* // Труды 11й всероссийской научной конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – Петрозаводск, 2009. – С. 69–77.
11. ТУЗОВСКИЙ А.Ф. *Онтолого-семантические модели в корпоративных системах управления знаниями:* дис. докт. техн. наук. – Томск, 2007. – С. 175–182.
12. ЧЕРНИЙ А.В., ТУЗОВСКИЙ А.Ф. *Развитие информационной системы организации с использованием семантических технологий* // Материалы Всерос. конф. с междунар. участием «Знания – Онтологии – Теория». – Новосибирск: ЗАО «РИЦ Прайс-Курьер», 2009. – Т.2. – С. 52–59.

13. ANDREASEN T., KNAPPE R., BULSKOV H. *Domain-specific similarity and retrieval* // 11th Int. Fuzzy Systems Association World Congress. – Vol. 1. – P. 496–502.
14. BONDY J.A., MURTY U.S.R. *Graph Theory*. – N.Y.: Springer, 2008. – 651 p.
15. BULSKOV H., KNAPPE R., ANDREASEN T. *On measuring similarity for conceptual querying* // Proc. 5th Int. FQAS Conf. LNCS. – Berlin: Springer, 2002. – Vol. 2522. – P. 100–111.
16. CASTANO S., FERRARA A., MONTANELLI S., RACCA G. *Semantic information interoperability in open networked systems* // Proc. of the Int. Conf. SNW. – Paris, 2004. – P. 215–230.
17. CHRISTOPHER D.M., PRABHAKAR R., HINRICH S. *Introduction to information retrieval*. – N. Y.: Cambridge University Press, 2008. – 482 p.
18. EHRIG M., SURE Y. *Ontology mapping – an integrated approach* // The Semantic Web: Research and Applications. Proc. 1st European Semantic Web Symposium. – Berlin: Springer. – Vol. 3053. – P. 76–91.
19. HAASE P., SIEBES R., HARMELEN F. *Peer selection in peer-to-peer networks with semantic topologies* // Proc. of Int. Conf. on Semantics in a Networked World. – Paris, 2004. – P. 108–125.
20. HIRST G., ST-ONGE D. *Lexical chains as Representations of context for the detection and correction of Malapropisms* // WordNet: an Electronic Lexical Database. – Cambridge: The MIT Press, 1998. – P. 305–322.
21. JIANG J., CONRATH D. *Semantic similarity based on corpus statistics and lexical taxonomy* // Proc. Int. Conf. on Computational Linguistics. – Taiwan, 1997. – P. 19–33.
22. KLIR G.J., YUAN B. *Fuzzy sets and fuzzy logic : theory and applications*. – N. Y.: Prentice Hall PTR, – 1995. – 574 p.
23. KNAPPE R. *Measures of semantic similarity and relatedness for use in ontology-based information retrieval*. PhD thesis. – Roskilde University, 2006. – 143 p.

24. LEACOCK C., CHODOROW M. *Combining local context and wordnet similarity for word sense identification* // WordNet: An Electronic Lexical Database. – Massachusetts: MIT Press, 1998. – P. 265–283.
25. LEVENSHTEIN I.V. *Binary codes capable of correcting delation, insertion and revercals* // Cybernetics and Control Theory. – 1966. – Vol.10., №8. – P. 707–710.
26. LIN D. *An information-theoretic definition of similarity* // Proc. 15th Int. Conf. on Machine Learning. – Massachusetts: Morgan Kaufmann, 1998. – P. 296–304.
27. MAEDCHE A., ZACHARIAS V. *Clustering ontology-based metadata in the Semantic Web* // Proc. 6th European PKDD Conf. LNCS. – Berlin: Springer, 2002. – Vol. 2431. – P. 348–360.
28. MILLER G., CHARLES W. *Contextual correlates of semantic similarity* // Language and Cognitive Processes. – 1991. – Vol.6, №1. – P. 1–28.
29. NGUYEN H.A. *New semantic similarity techniques of concepts applied in the biomedical domain and wordnet* // Thesis for the Degree Master of Science. – University of Houston-Clear Lake, 2006. – 108 p
30. PENIN T., WANG H., TRAN T., YU Y. *Snippet generation for semantic web search engines* // Proc. of the 3rd Asian Semantic Web Conf. on the Semantic Web. – Berlin: Springer Verlag, 2008. – P. 493–507.
31. RADA R., MILI H., BICKNELL E. ET AL. *Development and application of a metric on semantic nets* // IEEE Transactions on Systems, Man, and Cybernetics – 1989. – Vol.19, №1. – P. 17–30.
32. RESNIK P. *Using information content to evaluate semantic similarity in a taxonomy* // Proc. 14th Int. Joint Conf. on Artificial Intelligence. – 1995. – P. 448–453.
33. RUBINSTEIN H., GOODENOUGH J. *Contextual correlates of synomymy* // Communications of the ACM. – 1965. – Vol.8, №10. – P. 627–633.

34. SHANNON C.E., WEAVER W. *A mathematical theory of communication* // ACM SIGMOBILE Mobile computing and communications review. – 2001. – Vol.5, №10. – P. 3–55.
35. SPASIC I. NENADIC G., MANIOS K., ANANIADOU S. *Supervised learning of term similarities* // Proc. 3rd Int. IDEAL Conf. LNCS. – Berlin: Springer, 2002. – Vol. 2412. – P. 429–434.
36. SUSSNA M. *Word sense disambiguation for free-text indexing using a massive semantic network* // Proc. 2nd Int. Conf. IKM. – N.Y.: ACM Press, 1993. – P. 67–74.
37. TVERSKY A. *Features of similarity* // Psychological Rev. – 1977. – Vol. 84. – P. 325–352.
38. ZHU H., ZHONG J., LI J., YU Y. *An approach for semantic search by matching RDF graphs* // Proc. LAIRS Conf. – 2002. – P. 450–454.
39. WU Z., PALMER M. *Verbs semantics and lexical selection* // Proc. 32nd ann. Meeting ACL. – NJ, USA, 1994. – P. 133–138.

INFORMATION RETRIEVAL MODEL BASED ON SEMANTIC METADATA

Ba Ngoc Nguyen, Tomsk Polytechnic University, Tomsk, PhD student (nguyen_bn@hotmail.com).
Anatoly Tuzovsky, Tomsk Polytechnic University, Tomsk, Doctor of Science, professor.

Abstract: Typically, existing ontology-based approaches to semantic search use semantic similarity between ontology concepts and individuals as basic building blocks. We propose a semantic similarity scheme for triples and for sets of triples. Based on the proposed concept of semantic similarity, a theoretical model of semantic search was proposed. The model uses presentation of documents and queries in the form of triple sets.

Keywords: ontology, semantic similarity, semantic search, semantic metadata, information retrieval model.

Статья представлена к публикации членом редакционной коллегии В. Г. Лебедевым